Navy Personnel Research, Studies, and Technology Division
Bureau of Naval Personnel (NPRST/PERS-1)

Millington, TN 38055-1000 NPRST-TN-06-9 October 2006

Investigation of Item-Pair Presentation and Construct Validity of the Navy Computer Adaptive Personality Scales (NCAPS)

Christina M. Underhill, Ph.D.

Approved for public release; distribution is unlimited.

NPRST
research at work

NPRST-TN-06-9
October 2006

Investigation of Item-Pair Presentation and Construct Validity of the Navy Computer Adaptive Personality Scales (NCAPS)

Christina M. Underhill, Ph.D.

Reviewed and Approved by
Jacqueline A. Mottern, Ph.D.
Institute for Selection and Classification

Released by
David L. Alderton, Ph.D.
Director

Navy Personnel Research, Studies, and Technology (NPRST/PERS-1)
Bureau of Naval Personnel
5720 Integrity Drive
Millington, TN 38055-1000
www.nprst.navy.mil

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)* 30-09-2006	2. REPORT TYPE Technical Note	3. DATES COVERED *(From - To)* January 2004 through August 2004

4. TITLE AND SUBTITLE

Investigation of Item-Pair Presentation and Construct Validity of the Navy Computer Adaptive Personality Scales (NCAPS)

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

0602236N and 0603236N

6. AUTHOR(S)

Christina M. Underhill

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Navy Personnel Research, Studies, and Technology (NPRST/PERS-1)
Bureau of Naval Personnel
5720 Integrity Dr.
Millington, TN 38055-1000

8. PERFORMING ORGANIZATION REPORT NUMBER

NPRST-TN-06-9

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research
800 North Quincy Ave.
Ballston Tower One
Arlington, VA 22217-5660

10. SPONSOR/MONITOR'S ACRONYM(S)

ONR/Code 34

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

A - Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report documents one of the steps in our development of the Navy Computer Adaptive Personality Scales (NCAPS). NCAPS is a computer adaptive personality measure being developed and validated for use in the selection and classification of Sailors for entry level Navy enlisted jobs. This is an important component of our research program to overhaul and improve the Navy's enlisted selection and classification process. The over program, Whole Person Assessment, is designed to replace the current classification algorithm with a more flexible and accurate one that will also allow us to de-emphasize the almost exclusive focus on mental ability by including personality and interest measures in making classification decisions. NCAPS uses a cutting-edge technological approach to personality measurement which is designed to mitigate many problems that plague traditional instruments. Specifically, traditional instruments use Likert rating scales, and therefore are subject to both directed faking and socially desirable responding. To minimize these problems, NCAPS is developing a paired forced-choice item format, which uses a complex item response theory (IRT) adaptive selection and scoring algorithm. The complexity and novelty of the design constraints requires a series of interrelated research projects. This report is one in the series and fulfills the need to further explore the adaptive components and construct validity of NCAPS.

15. SUBJECT TERMS

Personality Assessment, Computer Adaptive Testing, Paired Comparison, Item-pair, Whole Person Assessment, NCAPS, Navy Computer Adaptive Personality Scales, selection and classification

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Genni Arledge
UNCLASS	UNCLASS	UNCLASS	UNCLASS	20	19b. TELEPHONE NUMBER *(Include area code)* 901-874-2115 (882)

Foreword

This report documents one of the steps in our development of the Navy Computer Adaptive Personality Scales (NCAPS). NCAPS is a computer adaptive personality measure being developed and validated for use in the selection and classification of Sailors for entry level Navy enlisted jobs. This is an important component of our research program to overhaul and improve the Navy's enlisted selection and classification process. The over program—Whole Person Assessment—is designed to replace the current classification algorithm with a more flexible and accurate one that will also allow us to de-emphasize the almost exclusive focus on mental ability by including personality and interest measures in making classification decisions. Collectively, these efforts would transform and modernize enlisted classification by making it applicant-centric while improving job satisfaction and performance, reducing attrition, and increasing continuation behavior.

NCAPS uses a cutting-edge technological approach to personality measurement which is designed to mitigate many problems that plague traditional instruments. Specifically, traditional instruments use straight-forward Likert rating scales, generally contain sets of homogeneous items, and therefore are subject to both directed faking and socially desirable responding. To minimize these problems, NCAPS is developing a paired forced-choice item format, uses a complex item response theory (IRT) adaptive selection and scoring algorithm, and intersperses item content. The complexity and novelty of the design constraints requires a series of interrelated research projects. This report is one in the series and fulfills the need to further explore the adaptive components of NCAPS along with additional construct validity assessment.

The research was sponsored by the Office of Navy Research (Code 34) and funded under PE 0602236N and PE 0603236N.

David L. Alderton, Ph.D.
Director

Executive Summary

This document details the results of an experiment to further investigate item presentation and construct validity of Navy Computer Adaptive Personality Scales (NCAPS) that, once fully-developed and validated, can be used by the Navy to improve selection and classification of Navy recruits. Currently, the Navy places recruits into jobs based on job availability and the recruit's Armed Services Vocational Aptitude Battery (ASVAB) scores (a cognitive ability measure). Individual preferences are only taken into consideration during a brief interview with a classifier. A recruit's personality is neither measured nor matched to jobs that may suit them best. Due in part relying almost solely on cognitive ability, over one-third of Sailors leave before they finish their first term of enlistment. A personality measure such as NCAPS can be used in conjunction with the ASVAB to improve job placement by creating a better person-job fit, enhancing job performance and satisfaction, thereby reducing attrition.

Objective

Results of a previous pilot test indicated that further investigation of the item selection method and construct validity of NCAPS was warranted (see Houston et al., 2003). The relationship as indicated by correlation coefficients between the adaptive and traditional version of NCAPS were not as strong as anticipated. In addition, almost all of the participants in the previous pilot test took the maximum number of item pairs allowed by the computer. This indicates that part of the adaptive mechanism for varying the number of items presented may not be working as efficiently as planned. For this study, we hypothesized that increasing the maximum number of item-pairs presented would yield stronger trait estimates from NCAPS by allowing more item pairs to be used in trait estimation.

Approach

Students from the University of Memphis served as participants in this study and were alternately assigned to take NCAPS with either a maximum of 10 item-pairs per construct or 25 item-pairs per construct. All participants also took a traditional version of NCAPS and a previously validated personality test. Class performance, overall academic performance, and cognitive ability scores of the participants were obtained and analyzed in relation to the three personality traits tested, achievement motivation, stress tolerance, and social orientation.

Results

- Trait scores obtained by the adaptive and traditional NCAPS were all significantly related to scores obtained on a validated personality test, indicating NCAPS measured the intended constructs.

- As expected, the personality estimates were not significantly related to cognitive ability.
- Personality estimates from NCAPS predicted aspects of performance above what can be explained by cognitive ability alone.
- Traits measured by NCAPS were predictive of class and overall academic performance indicating NCAPS' potential for use in predicting performance during Naval training.
- Analyses of the differences between item-pairs indicate that further investigation about item-pair selection is warranted.
- The item cutoff adaptive component of the Adaptive NCAPS version did not meet expectations. Suggestions for remediation are provided.
- The presentation of 25 pairs of items per construct is not efficient. Further studies are needed to find the optimal number of pairs to present that provides the most information in the shortest amount of time.

Recommendations

While the program is adaptive in the sense that the trait values of the pairs presented are dependent on an individual's previous answers, the number of item-pairs presented is not individually tailored and therefore all participants are required to take the same number of items. It is recommended that a larger scale validation project be done which explores (1) the utility of developing an algorithm that cuts off item presentation once participants reach asymptote and (2) the utility of presenting all participants a specifically predetermined number of items.

Contents

List of Tables

List of Figures

Introduction

Purpose

The purpose of the study was to further investigate the item-pair presentation process and its impact on the construct validity of NCAPS, a computer adaptive personality measure. Once fully developed and validated, NCAPS can be included in the Whole Person Assessment approach to improving selection and classification of recruits in the Navy. This study followed a beta test of NCAPS that assessed the feasibility of measuring personality traits reliably using state-of-the-art technology. This proof-of-concept study examined the construct validity of the traits measured using NCAPS compared to industry-accepted standards in traditional formats. Results from that pilot test indicated that NCAPS was measuring the traits intended, but that the relationships were not as strong as anticipated (Houston et al., 2003). The objective of this study is to examine the adaptive components of NCAPS by increasing the maximum number of item-pairs presented per construct.

Problem

The current Navy system for classifying new recruits for training programs and career tracks involves matching a recruit's ASVAB qualification score to the immediate needs of the Navy. The ASVAB measures cognitive ability and four specific abilities (i.e., verbal, numerical, technical, and perceptual speed). Based on scores averaged across tests, the new recruit is assigned to a training school and ultimately a Navy rating. Once recruits are assigned to a technical school and rating, there is very little opportunity to switch careers. The Navy does not utilize a process or measure that matches recruits' individual interests, preferences, or personality with available occupations. Recruits have very little input into which career path they are placed (Ferstl et al., 2003). The goal of the study is to develop a psychometrically sound personality assessment tool that can be used in conjunction with the ASVAB to better classify Sailors for jobs. This would improve person-job fit, and ultimately increase job performance, decrease attrition, and enhance job and career satisfaction.

Cognitive ability measures such as the ASVAB are generally good at predicting whether or not a recruit will successfully complete his or her training program. Once a Sailor progresses on to his or her job, other factors such as work related attitudes, which are influenced by personality, determine successful job performance. In their review Borman et al. (2003) note that examples of individual difference variables that contribute to overall job performance include person-organization and person-job fit, and attributes such as conscientiousness, emotional stability, extroversion, sociability, personal adaptability, integrity, and strategic career orientation. Being able to better match a recruit's abilities and personality with the needs of the Navy should result in a Sailor who fits better with his or her job and the Navy. This improvement in job classification should lead to a more satisfied recruit who will perform better on the job and be more likely to finish his or her first term of enlistment and reenlist (Borman et al., 2003).

Role of Personality in Selection

The goal of researchers in personnel selection and classification is to develop measures that predict future job performance. Employers want employees to not only perform well on the job but also remain on the job. Measures given to job applicants need to assess the knowledge, skills, and abilities necessary for successful performance of a particular job, ideally without producing adverse impact on subgroups of people. Cognitive ability is by far the best predictor of both training and job performance (Hunter & Hunter, 1984; Ree, Earles, & Teachout, 1994; Schmidt & Hunter, 1998). Cognitive ability can predict who will be a successful performer, but it is not sufficient for predicting whether a person will fit well with his or her organization and remain on the job.

Research has shown that one’s personality, motivation, and interest can substantially help predict turnover, retention, and job performance (Borman et al., 2003). Cognitive ability predicts knowledge components of job performance, whereas personality variables are better at predicting motivational components of performance (McCloy, Campbell, & Cudek, 1994), which influence turnover and retention. Employers, such as the U.S. Navy, who spend a great amount of time and money on training new employees or “recruits” can benefit from additional measures that better match an individual to a job.

In Schmidt and Hunter’s (1998) review of 85 years of selection methods in personnel psychology, they found that cognitive ability was the most valid predictor of training success (r = .56) and job performance (r = .51). When integrity and conscientiousness tests were added to cognitive ability, they provided incremental validity of .14 and .09, respectively, in predicting job performance (Schmidt & Hunter, 1998). Meta-analyses by McDaniel, Morgeson, Finnegan, Campion, and Braverman (2001) and Schmidt and Hunter (1998) found that conscientiousness—a personality trait—and situational job tests can improve performance prediction by 18 percent when used with cognitive ability. Interest inventories and biodata instruments, on the other hand, can improve prediction by only 2 percent in addition to cognitive ability (Schmidt & Hunter, 1998).

A measure with the most potential to provide the incremental validity beyond cognitive ability is one that is not highly correlated with cognitive ability. Personality measures have been shown to have little or no relation to cognitive ability (Ackerman & Heggestad, 1997; Day & Silverman, 1989; McHenry, Hough, Toquam, Hanson, & Ashworth, 1990). Interest inventories, on the other hand, are thought of as weakly correlated to cognitive ability (Ackerman & Heggestad, 1997). Biodata measures and situational job tests have been found to be moderately correlated to cognitive ability, but the strength of the relationship varies across different scales and tests (Allworth & Hesketh, 1999; Schmidt, 1988). Studies by Borman, White, and Dorsey and by Borman, White, Pulakos, and Oppler (as cited in Ferstl et al., 2003), found that the variance accounted for in job performance can increase substantially when personality measures are used in conjunction with cognitive ability measures.

Ferstl et al. (2003) have also cited research that personality measures produce the least amount of subgroup differences. Cognitive ability tests produce differences between black and white test takers more than any other measure (Hunter & Hunter, 1984). Situational job tests and biodata instruments produce less racial differences, but they are still not as good as personality measures in minimizing racial or gender differences (Borman et al., 2003). A non-cognitive measure is better at reducing adverse impact on race, gender and age subgroups (Hough, Oswald, & Ployhart, 2001).

Computer Adaptive Technology

The main principle behind adaptive ability testing used in employee selection is that the person's responses are used to modify the test while he or she is in the process of taking it. The test is modified so that the criterion used to estimate a person's ability is reached as efficiently as possible. One method of adaptive item presentation is based on item difficulty. If a participant responds correctly, then he or she is presented with a harder item. If the participant responds incorrectly, he or she is presented with an easier item. Items are presented until the participant consistently answers items correctly at a specific level of difficulty, at which point he or she is not presented with any more items (Bartram, 1993).

In many testing environments, including military personnel testing, there is a limited amount of time available for assessment. Therefore the purpose of computer adaptive testing is to present items that are informative about the test taker, and to maximize the precision of measurement in a limited amount of testing time. For example, if a high ability person receives all the same easy items as everyone else, his or her ability could not be accurately measured until he or she answers the harder items. The more difficult items better distinguish his or her ability from someone with lesser ability who would get those items wrong. In a computer adaptive test, items presented to examinees would more closely approximate their ability level so they would not have to answer too many items above or below their ability level. Thus, the item presentation would provide a more accurate measure of a test taker's ability (Wainer & Mislevy, 2000).

Bartram (1993) identified several advantages to the use of computer adaptive testing. Computer adaptive testing can use fewer items to assess someone's ability thereby reducing test length and time to administer. Computer adaptive technology can be used to obtain good trait level estimates. Computer adaptive technology can also provide better differentiation between participants' ability because of its capability to represent a wider range of difficulties within one instrument. Compared to paper and pencil tests, computer adaptive tests have good reliability, and they can be scored almost instantly.

NCAPS Development

Computer adaptive tests that have been developed since the invention of high-powered, inexpensive computing (e.g., Graduate Record Examination [GRE] and American College Test [ACT]) test job knowledge and cognitive ability. Computer adaptive technology (CAT) has not yet been applied to the measurement of personality; therefore, there is very little research regarding computer adaptive personality testing (Ferstl et al., 2003; Wainer et al., 2000). Prior to NCAPS, there have been no reports of a functional computer adaptive personality measure in the literature. When measuring personality as opposed to measuring cognitive ability, there is no right or wrong answer or degree of difficulty associated with the items in the measure. Items on a personality measure are differentiated by how much each statement represents a particular personality trait. For example, a statement representing someone with low achievement would read, “I only take on projects that I expect will be easy to complete.” A statement representing someone with high achievement would read, “I usually set difficult goals for myself.” For a complete description of item development and trait scaling please see Ferstl et al. (2003) and Houston et al. (2005).

The NCAPS is a paired comparison forced choice measure. Paired comparison forced choice formats have been shown in other studies to be resistant to intentional response distortion (Jackson, Wroblewski, & Ashton, 2000; Martin, Bowen, & Hunt, 2002). NCAPS administers personality statements in pairs, with each pair representing the same personality trait. During testing, the statement “I always do the work that is expected of me” (rated a 3) could be presented with the statement “I like to set goals that force me to perform at a higher level than what I’ve done in the past” (rated a 4). Test participants are forced to choose the statement that best represents them.

One adaptive component of NCAPS is that the computer algorithm selects the next pair of items to present according to the trait value of the last item selected. The trait values of the next pair for that construct would essentially bracket the trait value of the last item endorsed. If the previous item selected had a trait value of 3, then in the next pair of items presented, one statement would have a trait value slightly higher than 3 and one statement would be slightly lower than 3. Item presentation and selection would continue in this manner until the variation of the trait values of the items selected by the participant becomes minimal, thereby enabling an automatic item cut-off such as found in adaptive ability testing. This second adaptive mechanism of having such an automatic cut off would allow for the number of items presented to participants to be individually tailored, thereby decreasing testing time overall and increasing efficiency.

The initial development of NCAPS was limited to three personality constructs: achievement motivation, stress tolerance, and social orientation. Since this was the first earnest attempt to apply computer adaptive technology to personality measurement, researchers wanted to make sure the program worked before developing scales for additional traits. For a full review and description of the three initial traits see Ferstl et al. (2003).

Achievement Motivation

Achievement is defined and used in NCAPS as a person’s motivation to set and achieve challenging goals, work hard, and persist in the face of significant obstacles. In their review of the literature, Ferstl et al. (2003) noted that in relation to the Big Five, achievement has been considered a facet of conscientiousness by many personality researchers. Conscientiousness has been found to be the best personality predictor of performance across a wide range of occupations (Barrick, Mount, & Judge, 2001). Studies by Salgado as well as by Schmidt and Hunter (as cited in Ferstl et al., 2003) found that measurement of achievement produces gains in incremental predictive ability of 11–18 percent over measures of cognitive ability alone. In a study of military personnel by Hough, Eaton, Dunnette, Kamp, and McCloy (as cited in Ferstl et al., 2003), achievement predicted effort and leadership, personal discipline, physical fitness and military bearing.

Stress Tolerance

Stress tolerance is defined as a person’s ability to maintain composure and think clearly under stressful situations. In supporting a measure of stress tolerance for inclusion in the initial development of NCAPS, Ferstl et al. (2003) cited information from studies by Barrick, Mount, and Judge as well as Judge and Bono noting that stress tolerance is considered to be part of emotional stability in the Big Five model of personality. Emotional stability has been found to be the next best predictor of successful job performance. A meta-analysis of military and civilian studies found that emotional stability could predict 10 percent additional variance in performance over cognitive ability alone. Emotional stability may be a greater factor in military performance than in civilian job performance. A study by Salgado reported that when the military studies were analyzed alone, emotional stability could predict an additional 38 percent of the variance in job performance (as cited in Ferstl et al., 2003).

Social Orientation

A person’s social orientation is the degree to which he or she likes to work alone or with others, whether he or she likes and readily accepts people, and how much he or she values connections with others. NCAPS developers chose facets of extroversion and agreeableness, both included in the Big Five model of personality, to define the social orientation construct. Extroversion and agreeableness seemed most relevant to Navy enlisted ratings such as Navy Counselor or Hospital Corpsman, where the ability to relate well to others and willingness to help others is important to the job (Ferstl et al., 2003). Each of the components of social orientation, extroversion, and agreeableness have been found to be predictive for different types of jobs. Extroversion is a better predictor of job performance in jobs that require high contact with people (Barrick et al., 2001), and Hough and colleagues found that agreeableness is the best predictor when evaluating performance in the teamwork aspects of jobs (as cited in Ferstl et al., 2003).

Proof-of-Concept Pilot Test

A pilot test was conducted with the first version of NCAPS, and the results were reported in Houston et al. (2003). Reserve Officer Training Corps (ROTC) students from two different universities took (1) NCAPS, (2) a traditonal (non-adaptive) version of NCAPS, and (3) a marker test. NCAPS was set to present a maximum of 10 item-pairs for each construct (Achievement Motivation, Stress Tolerance, and Social Orientation). The traditional version of NCAPS was an 89-item subset of the full 280-item NCAPS item pool. These items were administered on paper and presented as single statements with Likert scale response options. The marker test consisted of 91 items that were selected from the International Personality Item Pool and 3 personality tests developed and validated by a government contractor in connection with previous projects (See Ferstl, 2003).

The construct validity of the traditional and adaptive forms of NCAPS was assessed by comparing the trait scores to those on the marker test. Researchers found that the traditional version of NCAPS was more closely related to the marker test than the adaptive version. The correlations between the traditional version and the marker test for each construct ranged from .81 to .88, whereas the correlations between the adaptive version and marker test ranged between .48 and .67.

Researchers conducting the pilot test of NCAPS postulated five possible explanations of why the correlations for the adaptive version were much lower than the traditional version. First, the traditional version and the marker test were both administered by paper and pencil while the adaptive version was administered by computer. Common method variance between the traditional version and the marker test could be related to the low correlation between the traditional and adaptive versions. Second, the adaptive version and the traditional version may be measuring slightly different constructs. Third, the computer algorithm may not be selecting items properly. Fourth, the adaptive version may be repeatedly administering a particular subset of items and that subset may not be overlapping with the items on the traditional version or marker test. Lastly, the adaptive version may be excluding entire facets of a particular construct scale, which would reduce the correlation between measures (Houston et al., 2003).

Current Hypotheses

As this project was initiated, NCAPS was conceptualized as having two adaptive components similar to those used in adaptive ability testing. One component was that trait values of the item pairs presented are dependent on the item selected in the previous pair. Another adaptive component was that a participant’s responses are used to adjust the number of item pairs needed for each construct, thereby enabling an automatic cut off. Thus, test lengths should vary by person according to the consistency of that participant’s responses. NCAPS was also programmed with a maximum cut off for those individuals who do not enable the automatic cutoff. In the first pilot test, the program was set to present no more than 10 item-pairs per construct.

This study was designed around the hypothesis that the number of item-pairs presented in the pilot study was not sufficient to provide an accurate measure of personality. A participant may require more than 10 pairs of items to accurately measure his or her trait levels. It is expected that by increasing the maximum number of item-pairs presented to 25 the reliability of the measure will increase as will the correlation between NCAPS and validated personality measures. With more than 10 item-pairs there will be a greater opportunity for the computer algorithm to narrow in on a trait score before the maximum cutoff point is reached. It was expected that most participants will not need to take the maximum number of item-pairs, and that NCAPS will be more efficient in measuring personality traits than traditional format personality measures.

Method

Participants

Undergraduate students taking psychology courses were solicited to participate in the study. In exchange for their participation, the student's instructors offered them extra credit toward their course grade. Students were able to choose among 27 time slots over an 8-day period. A maximum of 10 participants could be tested during each time slot. A total of 134 students, 67 percent female and 33 percent male, participated in the study. The ages of the participants ranged from 18 to 53 with 77 percent of the sample between the ages of 18 and 21. The ethnic distribution was 55 percent Caucasian, 40 percent African American, 3 percent Asian or Pacific Islander, and the remaining 2 percent Hispanic. Forty percent of the students were freshmen, 28 percent were sophomores, 13 percent were juniors, 17 percent were seniors, and 2 percent indicated that they were "other."

Procedures and Measures

This study compared two groups of participants; one group was administered a version of NCAPS that presented a maximum of 10 item-pairs per construct, and the other group was administered a version that presented a maximum of 25 item-pairs per construct. Students were alternately assigned to the 10-item NCAPS or the 25-item NCAPS condition in the order that they arrived to take the tests. Participants were also asked to give permission for researchers to obtain their grade point average (GPA) and ACT scores from university records. GPA and class grades were used as a measure of performance, and ACT scores were used as an indicator of cognitive ability.

All participants were tested in the same room. Ten laptop computers provided by the Navy were set up around a conference room table, five administered the 10-item-pair condition and 5 administered the 25-item-pair condition. After completing the preliminary forms, participants began the NCAPS test followed by the marker test and the traditional NCAPS, all via computer. When each participant finished, his or her file was saved by the administrator who then provided the debriefing form. The entire testing protocol lasted between 45–60 minutes.

Participants in both groups were also administered the traditional form of NCAPS, the marker personality test, and a brief demographic questionnaire. The traditional form of NCAPS consisted of 88 items taken from the NCAPS item pool. Participants were presented with an item and asked to rate how much they agreed or disagreed with the statement on a 5-point Likert scale. The marker personality test consisted of items taken from the International Personality Item Pool and personality measures developed and validated by the contractors who conducted the pilot test (Houston et al., 2003). Results from the pilot test led researchers to drop items from the NCAPS item pool because of low item-scale correlations and reliabilities of items in the traditional NCAPS measure therefore only 94 items were administered in the marker test. All measures were administered in a computerized format that enabled the responses of the participants to be recorded directly to a database on each computer.

Random Response Check

There were five random response checks throughout the traditional NCAPS and marker test for which participants were asked to mark a certain response. Participants who responded incorrectly were identified. Those who marked two or more response checks incorrectly were considered to be randomly responding thereby making their responses invalid. Three participants met this criterion and their responses were eliminated from further analyses.

Data Scoring

Marker Test

Items on the marker test were scored from 1 "Very Inaccurate" to 5 "Very Accurate" or 1 "Definitely False" to 5 "Definitely True." All negatively worded items were reverse coded so that a larger number indicated a more positive trait. Trait scores for achievement, social orientation, and stress tolerance were computed by averaging the participant's responses for items of each trait.

Traditional NCAPS

The items on the traditional NCAPS were scored differently from the marker items. These items were taken from the NCAPS item pool that held items representing trait levels along a 1 to 7 scale. Items in the marker test represented traits at the extreme ends of a scale (e.g., "I always do my best"). Responses from strongly agree to strongly disagree were equally weighted because all the statements have the same trait value. Items in the NCAPS pool had different trait values. These items represented different levels of a trait rather than an extreme end of a scale like those in the marker test. For example, someone's response "strongly agree" to an item that is rated 3 (e.g., "I try to do my best at most things") is not equivalent to his or her response "strongly agree" to an item representing a 7 (e.g., "I excel at virtually everything I try").

The traditional NCAPS constructs were scored by the same method of scoring used in the pilot test. Computations were made to standardize responses based on each item's trait level and a person's response to that item. Table 1 was reproduced from Ferstl et al. (2003) and shows the weights given to a participant's response for a particular item's trait level. The item trait levels in the table show whole numbers for example purposes, but the actual trait values could range from 1 to 7. Formulas were created for every possible response to every individual item in the traditional NCAPS. Once a standardized response was calculated for each person's response, items for each construct were then averaged to compute an overall trait score for each personality dimension per participant.

Table 1
Score values assigned to traditional NCAPS items, by trait level and response

Traditional NCAPS Response Scale	Item Trait Level						
	2	3	4	5	6	7	8
Strongly Disagree 1	6.0	5.0	4.0	3.0	2.0	1.0	0.0
Disagree 2	4.5	4.0	3.5	3.0	2.5	2.0	1.5
Neither Agree Nor Disagree 3	3.0	3.0	3.0	3.0	3.0	3.0	3.0
Agree 4	1.5	2.0	2.5	3.0	3.5	4.0	4.5
Strongly Agree 5	0.0	1.0	2.0	3.0	4.0	5.0	6.0

Note: Reproduced from Ferstl et al. (2003). *Following the roadmap: Evaluating potential predictors for Navy selection and classification* (Technical Report No. 421). Minneapolis, MN: Personnel Decisions Research Institutes.

Results

Group Comparisons

Independent samples t-tests were conducted to see if there were any significant differences between the 10-item-pair and 25-item-pair groups on construct scores for each measure. There were no significant differences between the two groups for any measure of the constructs. Means for the constructs measured by NCAPS were higher than the traditional or marker means because they were measured on a different scale. NCAPS constructs were scored on a scale of 1 to 7, while NCAPS traditional and marker test were scored on a scale of 1 to 5. Table 2 shows the means for each group on each of the three personality constructs measured by NCAPS, traditional NCAPS, and the marker test.

Table 2
Group comparisons by measure

Measure	Group	*n*	*M*	*SD*	M Diff.[a]	*F*
Achievement						
NCAPS	10-item	66	5.44	.82	.03	.10
	25-item	65	5.41	.89		
Traditional NCAPS	10-item	65	3.35	.51	.01	.17
	25-item	65	3.35	.48		
Marker Test	10-item	66	3.77	.56	.04	.99
	25-item	65	3.73	.59		
Social Orientation						
NCAPS	10-item	66	5.45	.81	-.03	.03
	25-item	65	5.48	.84		
Traditional NCAPS	10-item	65	3.43	.43	.05	.02
	25-item	65	3.39	.44		
Marker Test	10-item	66	3.81	.47	.09	.002
	25-item	65	3.72	.48		
Stress Tolerance						
NCAPS	10-item	66	5.17	1.00	-.11	1.23
	25-item	65	5.28	1.14		
Traditional NCAPS	10-item	65	3.17	.61	.01	2.38
	25-item	65	3.16	.71		
Marker Test	10-item	66	3.31	.53	.01	3.01
	25-item	65	3.30	.66		

Note: Scores on NCAPS were calculated on a scale of 1 to 7. Scores on the traditional NCAPS and Marker Test were calculated on a 1 to 5 scale.

[a] Mean difference

Scale Reliability

Because there were no significant differences between the two experimental conditions on construct means, they were all combined to perform scale reliability analyses. Scale reliability was conducted on the traditional NCAPS and the marker test. All item-scale correlations were sufficient and no items were dropped from the analyses. Constructs measured by the traditional NCAPS and marker test all had good reliability with alphas .84 or above. See Table 3 for the alpha coefficients as well as the number of items in each scale.

Table 3
Descriptive statistics and reliability for NCAPS, traditional NCAPS, and marker test

Measure	*M*	*SD*	Alpha	# of items
NCAPS (*n* = 131)				
Achievement	5.43	.85	n\a	n\a
Social Orientation	5.46	.83	n\a	n\a
Stress Tolerance	5.22	1.07	n\a	n\a
Traditional NCAPS (*n* = 130)				
Achievement	3.35	.49	.84	19
Social Orientation	3.40	.43	.86	36
Stress Tolerance	3.17	.66	.89	23
Marker Test (*n* = 131)				
Achievement	3.75	.57	.88	20
Social Orientation	3.77	.48	.90	34
Stress Tolerance	3.31	.59	.93	37

Because participants did not receive all the same items on the NCAPS measure, internal consistency reliabilities could not be obtained as they were for the traditional NCAPS or marker test. The reliability of NCAPS was assessed in a small separate study conducted on a group of 21 researchers and support staff. In this study, each participant took the 25-item version of NCAPS twice. The scores for each construct were correlated to estimate reliability of scores over administrations. The stress tolerance construct scores were the most highly correlated with an alpha of .915. Achievement scores were correlated at .828, and social orientation scores were correlated at .766.

Construct Validity

Adaptive and Traditional

To judge whether or not the NCAPS and traditional NCAPS were measuring the same constructs as the marker test, correlations were conducted. A strong correlation between the two measures indicates that they are measuring the same constructs. When comparing construct scores on the NCAPS and traditional NCAPS, results indicated that there were stronger correlations among the 25-item-pair condition than among the 10-item-pair condition. All correlations between NCAPS and traditional NCAPS measures of the same construct were significant. The correlations between scores from the 25-item-pair NCAPS and the traditional NCAPS ranged from .754 to .820. In the 10-item-pair group, the correlations were again significant but slightly lower and ranged from .572 to .703 (see Table 4). One would expect that the correlations between the NCAPS and traditional NCAPS to be strong and positive because the items from the traditional NCAPS came from the NCAPS database. It is not unexpected for there to be a less than perfect correlation because the traditional NCAPS scores are based on only on a sample of the NCAPS items.

Table 4
Correlations between NCAPS and traditional NCAPS

	Traditional NCAPS		
NCAPS (adaptive)	**Achievement**	**Social Orientation**	**Stress Tolerance**
	10 item-pair (*n* = 65)		
Achievement	**.613****	.001	.421**
Social Orientation	.018	**.572****	.410**
Stress Tolerance	.142	.259*	**.703****
	25 item-pair (*n* = 65)		
Achievement	**.754****	.506**	.458**
Social Orientation	.343**	**.820****	.454**
Stress Tolerance	.342**	.469**	**.807****

$^{*}p < .05$; $^{**}p < .01$.

Adaptive and Marker

High correlations between NCAPS and the marker test indicate that NCAPS is measuring similar constructs as the marker test. As shown in Table 5, the personality construct scores of the people in the 25-item-pair group were more strongly correlated with the scores on the same construct as measured by the marker test than scores from the 10-item group NCAPS. Correlations between measures of the same construct for the 10-item-pair group ranged from .597 to .690. The correlations of the 25-item-pair group had higher correlations that ranged from .749 to .818. The correlation between measures of the same construct on NCAPS and the marker test were significant for both groups, although the correlations were higher for people in the 25-item-pair version of NCAPS.

Table 5
Correlations between NCAPS and marker test

	Marker Test		
NCAPS (adaptive)	Achievement	Social Orientation	Stress Tolerance
		10 item-pair (*n* = 66)	
Achievement	**.690****	-.036	.396**
Social Orientation	.191	**.597****	.357**
Stress Tolerance	.185	.264*	**.665****
		25 item-pair (*n* = 65)	
Achievement	**.749****	.410**	.450**
Social Orientation	.409**	**.814****	.420**
Stress Tolerance	.395**	.396**	**.818****

$^{*}p < .05$; $^{**}p < .01$.

Traditional and Marker

The correlations between the traditional NCAPS and the marker test were higher than the ones between NCAPS/marker test and NCAPS/traditional NCAPS. Traditional NCAPS and the marker test had strong correlations even with scores from participants in the 10-item-pair group, which had consistently weaker correlations than the 25-item-pair group. In the 10-item-pair group, correlations between measures of the same construct ranged from .796 to .919. Correlations in the 25-item-pair group ranged from .809 to .901. Results indicate that the traditional NCAPS is measuring the same constructs as measured by the marker test (see Table 6).

Table 6
Correlations between traditional NCAPS and marker test

NCAPS (traditional)	Marker Test		
	Achievement	Social Orientation	Stress Tolerance
		10 item-pair (n = 65)	
Achievement	**.800****	-.003	.192
Social Orientation	.262*	**.819****	.326*
Stress Tolerance	.419**	.357**	**.796****
		25 item-pair (n = 65)	
Achievement	**.809****	.407**	.319**
Social Orientation	.504**	**.857****	.478**
Stress Tolerance	.417**	.484**	**.901****

*p < .05; **p < .01.

Discriminate Validity

To assess whether or not the constructs being measured are distinct from one another, the correlations among traits were examined for each measure. When distinct traits are being measured, then low intra-construct correlations between the different personality traits are expected. As shown in Table 7, the intra-construct correlations on the marker test were significant and ranged from .32 to .45. This pattern of significant relationships among the constructs was similar for the NCAPS and traditional NCAPS. The intra-construct correlations for traditional NCAPS ranged from .29 to .44, while the intra-construct correlations for NCAPS ranged from .26 to .49. Compared to both the marker test and the traditional NCAPS, the 10-item-pair NCAPS group had the lowest intra-construct correlations. These results indicate that the 10-item version of NCAPS more clearly distinguishes between the three personality constructs than the other measures used in this study.

Table 7
Intra-construct correlations

	Marker Test	Traditional NCAPS	NCAPS	
Constructs	**All** [a]	**All** [b]	**10-item-pair** [c]	**25-item-pair** [d]
AV – SO	.323**	.326**	.256*	.499**
AV – ST	.385**	.293**	.295*	.477**
SO – ST	.453**	.443**	.348**	.451**

[a] *n* = 131; [b] *n* = 130; [c, d] *n* = 65
*p < *.05; ***p < *.01.*

Item Selection Process

The item selection process of NCAPS entailed presenting a participant with subsequent pairs of items that bracket the trait value of the last item selected. Based upon discussions with the contractor developing the program, it was thought that the as the test progresses the bracket or distance between the two trait values of items presented should get smaller. In order to evaluate how items were being selected, the trait values for each pair of items presented for every participant were obtained. Next, the difference between the values of each pair of items for each construct was computed. The differences between the two trait values were averaged for each pair across participants, and then graphed. The figures show that the trait values of the item-pairs do not converge or get closer as pairs of items are presented. The difference between the trait values of the items presented actually increased with each pair (see Figures 1–3).

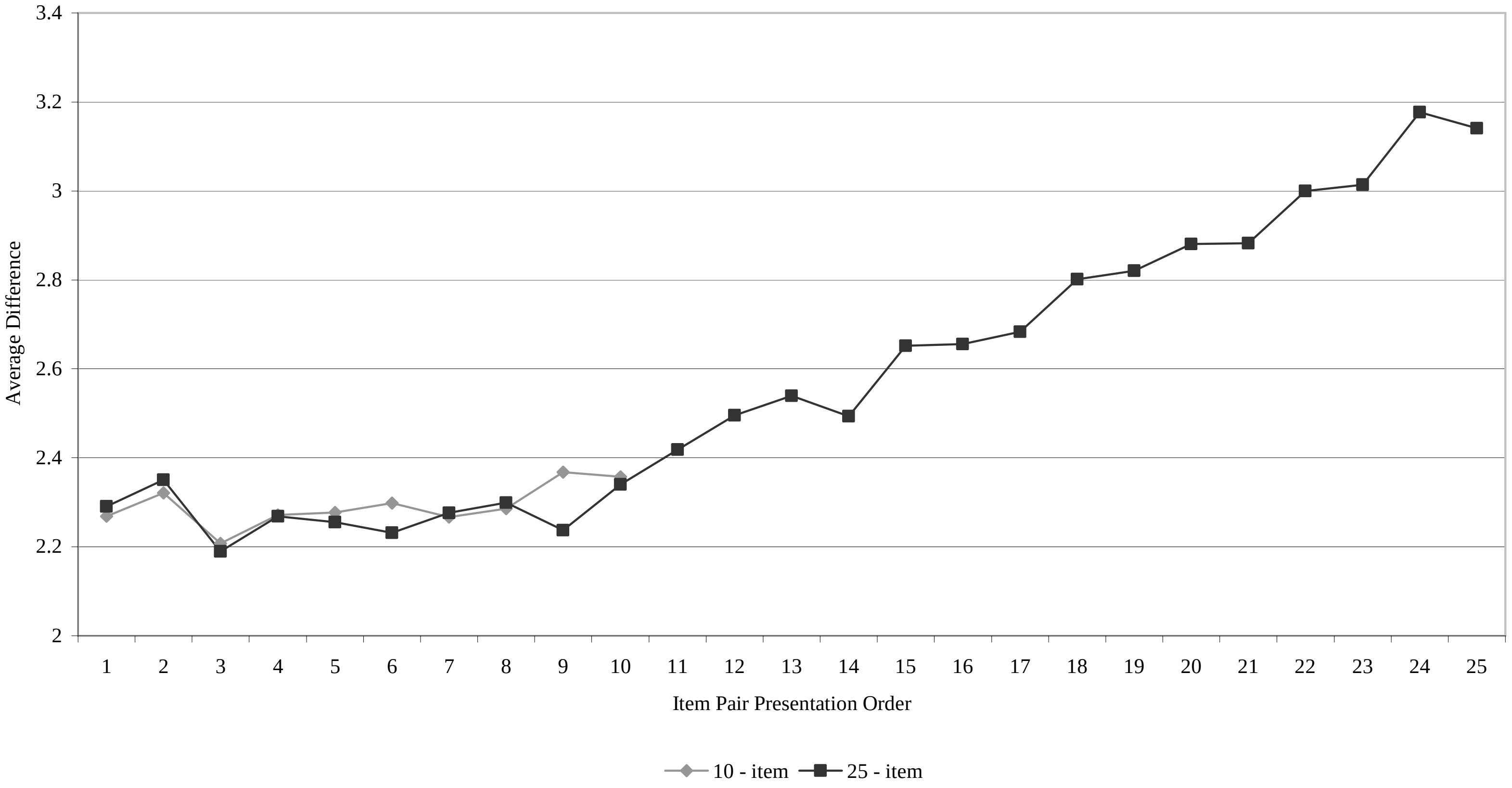

Figure 1. Difference between trait values of each pair of achievement items.

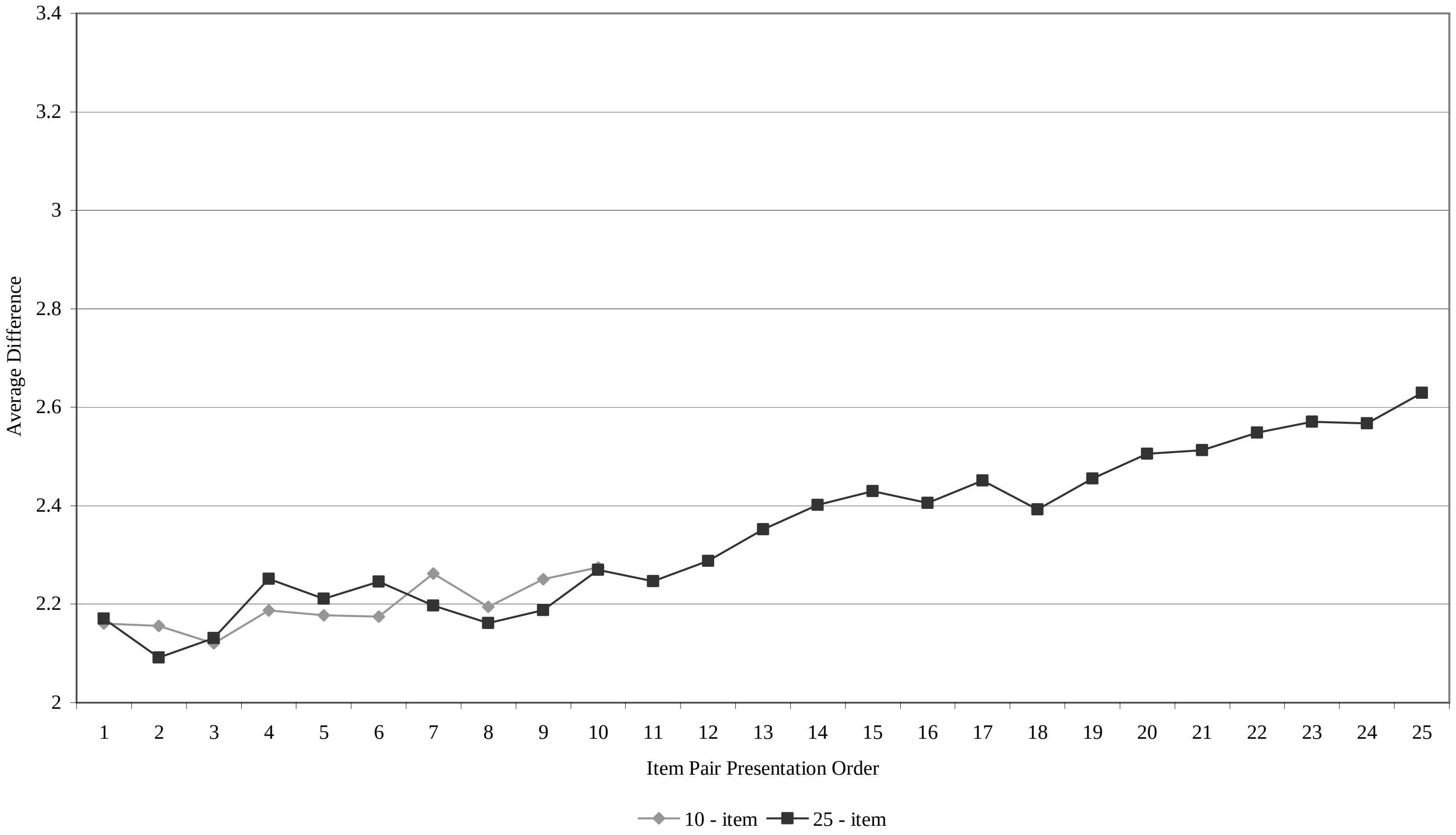

Figure 2. Difference between trait values of each pair of social orientation items.

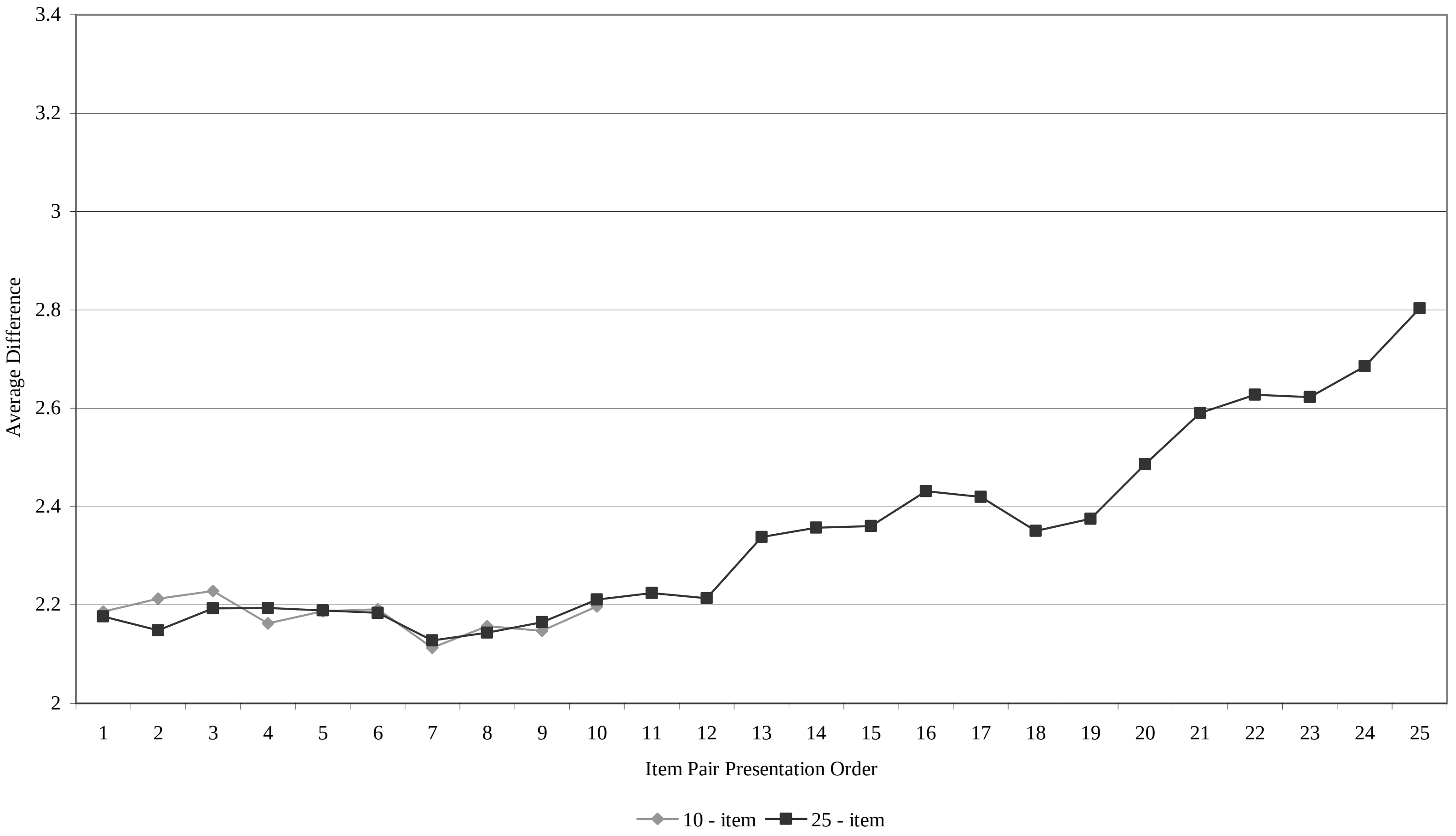

Figure 3. Difference between trait values of each pair of stress tolerance items.

For achievement, the average difference in trait values between the first pair of items was 2.28. The difference at the 10th pair increased to 2.35. At the 25th pair of items, the difference between the trait values had reached 3.14. For the social orientation and stress tolerance constructs, the trait value difference followed the same pattern of increasing distances as those of the achievement items. Between social orientation items the first pair difference was 2.17 and at the 25th pair the difference was 2.63. The range of differences in stress tolerance items was slightly larger than social orientation, starting at 2.18 and ending with a distance of 2.80. This trend could be caused because there were not enough items at the specific trait levels to allow for fine-grained differentiation, or it could be an indication that there is a problem with the item selection programming.

While the initial program manager believed that the program contained an adaptive component, which varied the number of items presented per participant based on his or her responses, this was not found to be the case. This is indicated by the number of participants who took the maximum number of pairs allowed by the program. Our hypothesis was that increasing the number to 25 would allow each participant enough pairs to answer so that the program could narrow in on a trait estimate and stop item presentation at different points for each test taker based on when the participant's answers reached an asymptotic state. Only a few people with erratic responding would ever take the max number allowed especially in the 25-item-pair version of NCAPS. In this study, 90 percent of the total participants took the maximum number of items allowed by the program in either of the two groups. See Table 8 for the number of item pairs taken and trait estimates for participants who took less than the maximum number of pairs allowed by the program.

Table 8
Participants who took less than the maximum number of item-pairs

Construct	Participant	# of items taken	Trait Estimate
Achievement			
	1	6	3.0284
	2	20	6.9572
	3	20	6.9782
	4	7	7.1681
	5	7	7.1754
Stress Tolerance			
	6	12	2.7560
	7	7	2.78
	8	6	2.8747
	9	6	2.8842
	10	6	2.8855
	11	18	3.1202
	12	19	7.0067
	13	15	7.0727
	14	8	7.3308

Note. All participants took the maximum number of social orientation item pairs.

Criterion Validity

The purpose of using NCAPS is to improve our ability to predict Navy recruits who will not only successfully complete their training program, but who will also be successful on the job. While it would be ideal to only utilize Navy recruits, university students were used for this initial study to preserve Recruits training time. Pending the results of this and other validation projects, the use of Navy Recruits will be planned. The relationship between the performance indicators (i.e., grade point average and class grades), cognitive ability indicator (i.e., ACT scores), and the three personality trait scores from NCAPS were evaluated using correlations and regression. Based on the performance and personality research, a strong correlation between participants' achievement motivation, GPA, and class grades was expected. Researchers also expected lower correlations between each personality trait and ACT scores as shown in Table 9.

Table 9
Performance and cognitive ability indicators correlated with personality traits as measured by NCAPS

	ACT	University GPA	High School GPA	Class Grades
ACT	1.0	-	-	-
University GPA	.228**	1.0	-	-
High School GPA	.400**	.447**	1.0	-
Class Grades	.491**	.311**	.466**	1.0
Achievement	-.007	.139	.206*	.339**
Social Orientation	-.069	.123	.008	.027
Stress Tolerance	-.116	-.120	-.048	-.085

Note: *$p < .05$: **$p < .01$.
Note: $n = 95$ to 108.

As expected, there were no significant relationships between the three NCAPS constructs and the composite ACT scores. The NCAPS traits were not related to cognitive ability. Only achievement motivation was significantly related to two of the performance indicators, high school GPA and class grades. Class grades as a performance indicator had the strongest relationship with ACT and achievement scores.

Adaptive NCAPS

In order to determine how well ACT and personality scores from NCAPS predict performance (e.g., GPA and class grades) stepwise regressions were conducted. ACT, achievement, social orientation, and stress tolerance were entered into regression models to predict each individual performance indicator. See Table 10 for model statistics of ACT and NCAPS scores as predictors of performance. Only ACT ($\Delta R^2 =$ .056) was found to be a significant predictor of university GPA. The personality constructs were excluded from the model because they did not meet the inclusion criteria of .09. ACT scores explained 5.6 percent of the variance in university GPAs, while personality scores accounted for no additional variance.

Table 10
Regression Models of ACT and NCAPS Constructs as Predictors of Performance

Performance Indicator	*df*	*F*	Model ΔR^2	*p*
University GPA [a]	1, 92	5.467	.056	.022
High School GPA [b]	2, 90	10.416	.188	.000
Class Grades [c]	2, 77	17.403	.317	.000

[a] ACT only significant predictor, ΔR^2 =.056
[b] ACT ΔR^2 = .154 and Achievement ΔR^2 = .034
[c] ACT ΔR^2 = .217 and Achievement ΔR^2 = .10

In the model predicting high school GPA, ACT (ΔR^2 = .154) and achievement scores (ΔR^2 = .034) were significant predictors. ACT explained 15 percent of the variance in high school GPA while achievement explained an additional 3.4 percent of the variance. When using class grades in a model as the performance indicator, ACT and achievement were again significant predictors. ACT (ΔR^2 = .217) and achievement (ΔR^2 = .10) together accounted for 31.7 percent of the variance in class grades. Of the three performance indicators, ACT and achievement motivation explained the most variance in class grades.

Traditional NCAPS

Stepwise regressions using personality scores from the traditional NCAPS constructs were also conducted. Traditional NCAPS scores had a stronger relationship with the marker test than adaptive NCAPS scores. Results from the regressions showed that ACT scores and traditional NCAPS achievement scores explained slightly more variance in some of the models than ACT and adaptive NCAPS achievement scores. See Table 11 for model statistics of ACT and traditional NCAPS scores as predictors of performance. None of the personality constructs as measured by the traditional NCAPS were significant predictors of university GPA. ACT (ΔR^2 = .154) and achievement (ΔR^2 = .056) were significant predictors of high school GPA explaining 21% of the variance. When predicting class grades, ACT (ΔR^2 = .217) and achievement (ΔR^2 = .138) were again the only significant predictors in the model, and they explained a total of 35.6 percent of the variance in class grades.

Table 11
Regression models of ACT and traditional NCAPS Constructs as predictors of performance

Performance Indicator	*df*	*F*	Model ΔR^2	*p*
University GPA [a]	1, 92	5.467	.056	.022
High School GPA [b]	2, 92	11.98	.21	.000
Class Grades [c]	2, 77	22.698	.356	.000

[a] ACT only significant predictor, ΔR^2 =.056.
[b] ACT ΔR^2 = .154 and Achievement ΔR^2 = .056.
[c] ACT ΔR^2 = .217 and Achievement ΔR^2 = .138.

Discussion

The main hypothesis of this study was that increasing the maximum number of pairs presented would allow for more accurate estimates of a person's personality trait. Trait estimates were evaluated on how the scores from the 25 item-pairs related to those trait estimates obtained from 10 item-pairs, and if the addition of 15 more items allowed the person's item selection to trigger the item presentation stopping rule *before* he or she reached the maximum number of items presented. Researchers found that the number of items presented per person does not vary based on their responses. However, the program is adaptive in that item-pairs presented to each person is dependent on that person's response to the previous pair of items. Because the program does not adjust the number of pairs presented, all participants took the maximum number allowed by the program. In this experiment participants took a version of NCAPS programmed to present either 10 item-pairs per trait or 25 item-pairs per trait. Further validation research is needed to find (1) the optimal number of pairs to present or (2) provide data to develop the algorithm to identify optimal item cut off after participants reach asymptote. Ideally, further work will identify the cutoff method, which will allow NCAPS to gain the most useful information about each participant in the shortest amount of time.

The trait scores calculated by both adaptive and traditional NCAPS were all significantly related to scores on the marker test. This indicated that NCAPS is measuring the traits intended. The magnitude of the relationships between NCAPS, traditional NCAPS, and the marker test were similar to the result from the first pilot test by Ferstl et al. (2003) indicating a consistency of measurement.

As expected based on the personality literature, findings of this study show that cognitive ability is not related to the personality traits measured by both NCAPS and traditional NCAPS. Results from the analyses show that achievement motivation accounted for unique variance in addition to that explained by cognitive ability. NCAPS achievement scores when used to predict class grades provided as much incremental validity as reported in the literature regarding achievement motivation.

It is not unexpected that stress tolerance and social orientation were not predictive of class performance. Previous research has shown that different personality constructs are more predictive for certain jobs because of the different skills and tasks required. Achievement is a better predictor of academic class performance because class performance is dependent on a person's motivation to achieve and their cognitive ability (e.g. capacity to learn). Achievement motivation is an influential trait or drive behind academic success. Success in jobs such as a fire fighter or police officer is partially dependent on their ability to think clearly and maintain composure under stressful situations. For jobs such as these, stress tolerance would be a better predictor than achievement motivation.

Investigation into the distance between the trait values of the item-pairs found that the distance did not decrease as expected, but rather increased. Contrary to our initial expectations, research by Stark and Drasgow (2002) found that when using a paired comparison IRT model on which NCAPS is based, the distance of the trait values of the pair of items do not need to decrease in order to obtain the maximum amount of information. In fact Stark and Drasgow found that a pair of statements with a distance of 2 between their values provided the maximum amount of information. This examination work was conducted because the increase in trait value distances beyond 2 was not an expected occurrence in this particular program. This work has confirmed that the program is indeed operating with the methodology suggested by Stark and Drasgow (2002).

The increase in distances in trait values was found to be a result of the NCAPS running out of items with particular trait values to present. The program limits the number of times the same item can be presented to each participant. Each item can be presented only twice. Referring back to Figures 1, 2, and 3, it appears that after 10 pairs the items at a particular trait level were exhausted. As the pool ran out of items with those values, it continued to present items with values more greatly separated, repeating the process until participants reached the end of the test. If there were more items in the item pool the computer would keep presenting items around the same values, therefore the line of trait value differences would go straight across the presentation order of items. More items in the pool would mean that the computer would not as frequently run out of items with a particular trait rating and have to move on to items with traits further away from each other.

Most participants took the maximum number of pairs presented in both groups, but there were a few participants who did not take the maximum. If the program was not designed to activate an automatic cut off that varied per participant, then how did some participants take less than the maximum programmed? As discussed previously, the program could only present the same item twice. The participants who took fewer items scored at an extreme end of the scale. When there were no more items with values higher or lower (depending on the end of the scale) to present to these people, the item presentation stopped and a score was calculated.

Overall Conclusion

Results indicated that NCAPS is predictive of class performance. While NCAPS was initially intended to predict performance in the fleet, it could be a useful tool predicting success through various Navy training programs. It is anticipated that NCAPS will have an important impact on the field of personality and the naval job classification procedures. By combining measures of personality with the ASVAB, the Navy will hopefully have a better-matched Sailor who will be more successful and satisfied with his or her job, and ultimately stay in the Navy. Recruiting and training costs can be reduced when more Sailors stay in the Navy. A computer adaptive measure that takes less time to administer and score, and produces more accurate estimates of personality traits would be ideal for the Navy.

Future Research or Research Consideration

Further validation research is needed to find either the optimal numbers of pairs to present per construct, or to provide data for developing the algorithm to identify optimal item cut off after participants reach asymptote. Large-scale validation of NCAPS has begun using training and fleet performance as criteria. These further studies will illuminate the utility of NCAPS as a classification device.

References

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin*, *121*, 219-245.

Allworth, E., & Hesketh, B. (1999). Construct-oriented biodata: Capturing change-related and contextually relevant future performance. *International Journal of Selection and Assessment*, *7*, 97-111.

Bartram, D. (1993). Emerging trends in computer-assisted assessment. In H. Schuler, J. L. Farr, & M. Smith (Eds.), *Personnel selection and assessment: Individual and organizational perspectives* (pp. 267-288). Hillsdale, NJ: Lawrence Erlbaum Assoc.

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9,* 52-69.

Borman, W. C., Hedge, J. W., Ferstl, K. L., Kaufman, J. D., Farmer, W. L., & Bearden, R. M. (2003). Current directions and issues in personnel selection and classification. In J. J. Martocchio & G. R. Gerris (Eds.) *Research in personnel and human resource management* (vol. 22). Amsterdam: Elsevier.

Day, D. V., & Silverman, S. B. (1989). Personality and job performance: Evidence of incremental validity. *Personnel Psychology*, *42*, 25-36.

Ferstl, K. L., Schneider, R. J., Hedge, J. W., Houston, J. S., Borman, W. C., & Farmer, W. L. (2003). *Following the roadmap: Evaluating potential predictors for Navy selection and classification* (Technical Report No. 421). Minneapolis, MN: Personnel Decisions Research Institutes.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection, and amelioration of adverse impact in personnel selection procedures: Issues, evidence, and lessons learned. *International Journal of Selection and Assessment, 9,* 152-194.

Houston, J. S., Borman, W. C., Farmer, W. L., & Bearden R. M. (2005). *Development of the Enlisted Computer Adaptive Personality Scales (NCAPS), renamed Navy Computer Adaptive Personality Scales (NCAPS)* (Technical Report No. 503). Minneapolis, MN: Personnel Decisions Research Institutes.

Houston, J. S., Schneider, R. J., Ferstl, K. L., Borman, W. C., Hedge, J. W., Farmer, W. L., & Bearden, R. M. (2003). *NCAPS: Development of the enlisted computer adaptive personality scales for the United States Navy* (Technical Report No. 449). Minneapolis, MN: Personnel Decisions Research Institutes.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*, 72-98.

Jackson, D. N., Wroblewski, V. R., & Ashton, M.C. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance, 13* (4), 371-388.

Martin, B. A., Bowen, C. C., & Hunt, S. T. (2002). How effective are people at faking on personality questionnaires? *Personality and Individual Differences, 32*, 247-256.

McCloy, R. A., Campbell, J. P., & Cudeck, R. (1994). A confirmatory test of a model of performance determinants. *Journal of Applied Psychology, 78*, 493-505.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*, 730-740.

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology, 43*, 335-354.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than g. *Journal of Applied Psychology, 79* (4), 518-524.

Schmidt, F. L. (1988). The problem of group differences in ability scores in employment selection. *Journal of Vocational Behavior, 33*, 272-292.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124* (2), 262-274.

Stark, S. & Drasgow, F. (2002). An EM approach to parameter estimation for the Zinnes and Griggs paired comparison IRT model. *Applied Psychological Measurement, 26* (2), 208-227.

Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration and proficiency estimation. In H. Wainer et al., *Computerized adaptive testing: a primer (*2nd ed., pp. 61-100). Mahwah, NJ: Lawrence Erlbaum Assoc.

Wainer, H., Dorans, N. J., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). Future challenges. In H. Wainer et al., *Computerized adaptive testing: a primer (*2nd ed., pp. 231- 269). Mahwah, NJ: Lawrence Erlbaum Assoc.

Distribution

AIR UNIVERSITY LIBRARY
ARMY MANAGEMENT STAFF COLLEGE LIBRARY
ARMY RESEARCH INSTITUTE LIBRARY
ARMY WAR COLLEGE LIBRARY
CENTER FOR NAVAL ANALYSES LIBRARY
DEFENSE TECHNICAL INFORMATION CENTER
HUMAN RESOURCES DIRECTORATE TECHNICAL LIBRARY
JOINT FORCES STAFF COLLEGE LIBRARY
MARINE CORPS UNIVERSITY LIBRARIES
NATIONAL DEFENSE UNIVERSITY LIBRARY
NAVAL HEALTH RESEARCH CENTER WILKINS BIOMEDICAL LIBRARY
NAVAL POSTGRADUATE SCHOOL DUDLEY KNOX LIBRARY
NAVAL RESEARCH LABORATORY RUTH HOOKER RESEARCH LIBRARY
NAVAL WAR COLLEGE LIBRARY
NAVY PERSONNEL RESEARCH, STUDIES, AND TECHNOLOGY SPISHOCK LIBRARY (3)
PENTAGON LIBRARY
USAF ACADEMY LIBRARY
US COAST GUARD ACADEMY LIBRARY
US MERCHANT MARINE ACADEMY BLAND LIBRARY
US MILITARY ACADEMY AT WEST POINT LIBRARY
US NAVAL ACADEMY NIMITZ LIBRARY